After fitting these linear models fit1 <- lm(bp ∼ obese,data=bp.obese), fit2 <- lm(bp ∼ obese+sex,data=bp.obese) and fit3 <- lm(bp ∼ obese\*sex,data=bp.obese), the following outputs are obtained by the R function summary.

##   
## Call:  
## lm(formula = bp ~ obese, data = bp.obese)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.570 -11.241 -2.400 9.116 71.390   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 96.818 8.920 10.86 < 2e-16 \*\*\*  
## obese 23.001 6.667 3.45 0.000822 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.28 on 100 degrees of freedom  
## Multiple R-squared: 0.1064, Adjusted R-squared: 0.09743   
## F-statistic: 11.9 on 1 and 100 DF, p-value: 0.0008222

fit2 <- lm(bp ~ obese+sex,data=bp.obese)  
summary(fit2)

##   
## Call:  
## lm(formula = bp ~ obese + sex, data = bp.obese)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.263 -11.613 -2.057 6.424 72.207   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 93.287 8.937 10.438 < 2e-16 \*\*\*  
## obese 29.038 7.172 4.049 0.000102 \*\*\*  
## sex -7.730 3.715 -2.081 0.040053 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17 on 99 degrees of freedom  
## Multiple R-squared: 0.1438, Adjusted R-squared: 0.1265   
## F-statistic: 8.314 on 2 and 99 DF, p-value: 0.0004596

fit3 <- lm(bp ~ obese\*sex,data=bp.obese)  
summary(fit3)

##   
## Call:  
## lm(formula = bp ~ obese \* sex, data = bp.obese)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -25.645 -11.621 -1.708 6.737 71.500   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 102.112 18.231 5.601 1.95e-07 \*\*\*  
## obese 21.646 15.118 1.432 0.155   
## sex -19.596 21.664 -0.905 0.368   
## obese:sex 9.558 17.191 0.556 0.579   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.05 on 98 degrees of freedom  
## Multiple R-squared: 0.1465, Adjusted R-squared: 0.1204   
## F-statistic: 5.607 on 3 and 98 DF, p-value: 0.001368

Describe how to interpret these results, and then suggest how to proceed with further analyses.

Sono tre modelli annidati di regressione su bp. Tutti e tre sono caratterizzati dai valori degli indici estremamente bassi (i modelli spiegano poco la varianza nella risposta).

Dal punto di vista della significatività dei parametri fit3 è il modello peggiore (tutti i regressori tranne l’intercetta hanno p-value elevato), in particolare l’interazione tra obese e sex; mentre il migliore è fit1.

Confrontiamo i tre modelli con ANOVA, AIC e BIC per scegliere il migliore tra fit1 e fit2

anova(fit1, fit2)

## Analysis of Variance Table  
##   
## Model 1: bp ~ obese  
## Model 2: bp ~ obese + sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 100 29846   
## 2 99 28595 1 1250.3 4.3289 0.04005 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(fit1, fit2)

## df AIC  
## fit1 3 874.7029  
## fit2 4 872.3376

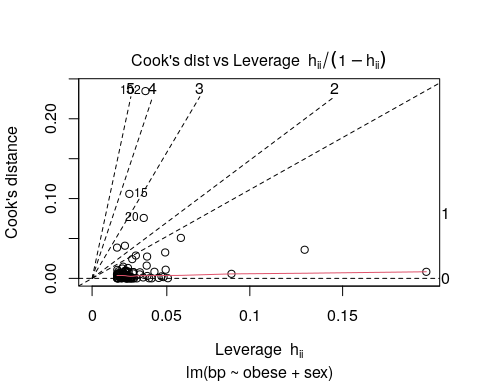
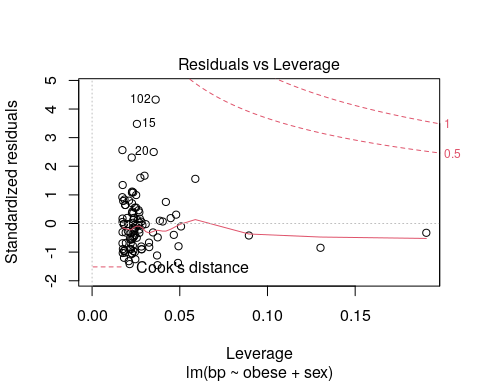
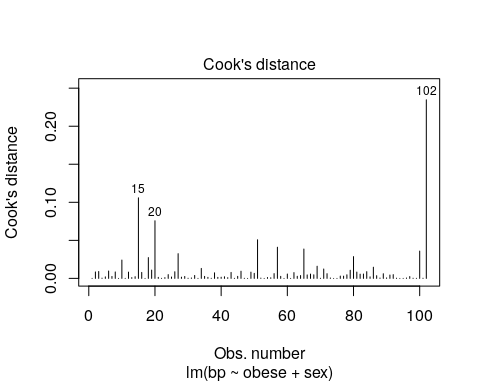
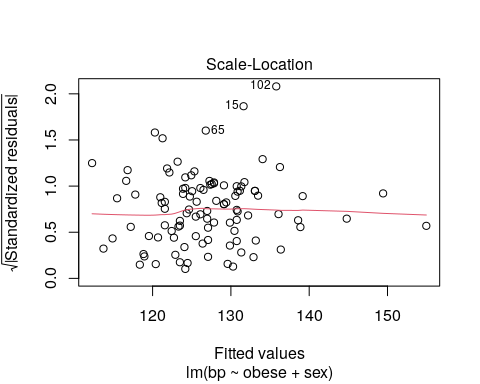
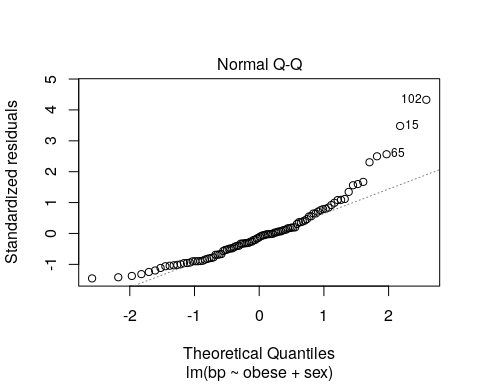
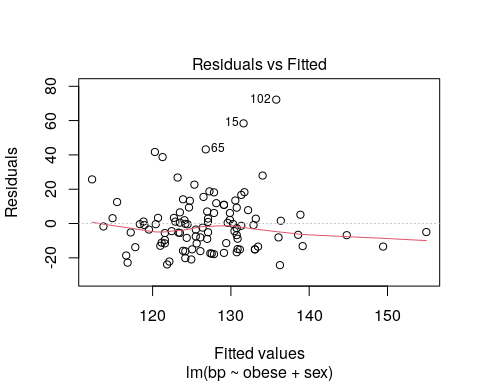
BIC(fit1, fit2)

## df BIC  
## fit1 3 882.5778  
## fit2 4 882.8375

L’analisi ANOVA e la statistica AIC suggeriscono di utilizzare il modello fit2, mentre la statistica BIC è sostanzialmente uguale per entrambe.

Scegliamo il modello fit2

plot(fit2, which = 1:6)



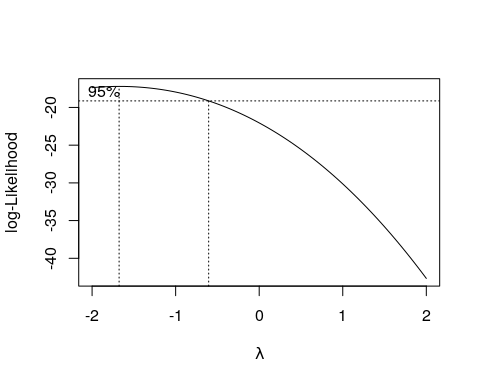
Dai grafici diagnostici del modello si notano molto bene gli outliers, i quali tuttavia sono i punti più influenti, ma la distanza di Cook e l’effetto leva non sono troppo elevati.

Risulta violata l’ipotesi di normalità dei residui dal grafico dei residui standardizzati, soprattutto sulle code.

Considerato che in fase di EDA si erano notate un alto livello di asimmetria e curtosi per le variabili numeriche, introduciamo delle trasformazioni.

Calcoliamo una trasformata adeguata per bp tramite il metodo di Box Cox.

library(MASS)  
boxcox(fit2)



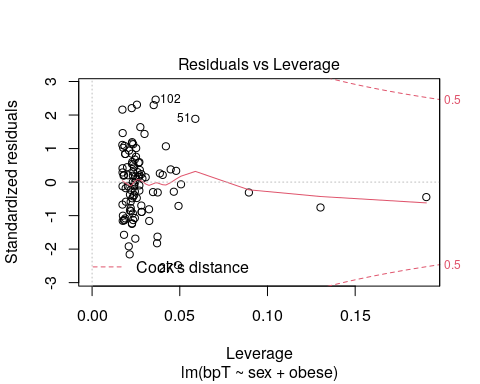
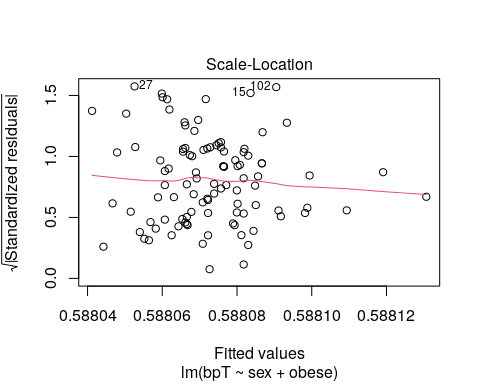
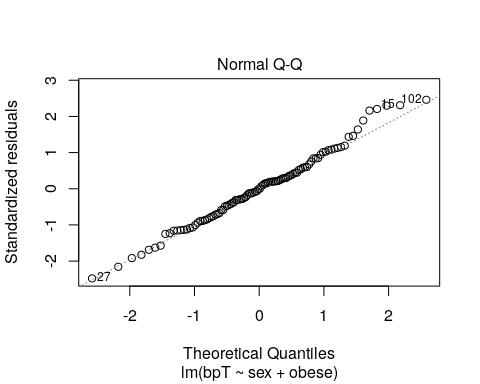
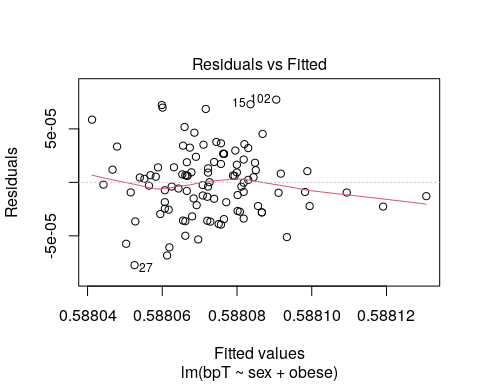
lambda <- boxcox(fit2, plotit = F)  
#Estrae il valore di x in corrispondenza in corrispondenza del valore massimo di y  
lambda.max = lambda$x[which.max(lambda$y)]  
print(lambda.max)

## [1] -1.7

bp.obese$bpT <- (bp.obese$bp^lambda.max - 1)/lambda.max  
fit4 <- lm(bpT~sex+obese, data = bp.obese)  
summary(fit4)

##   
## Call:  
## lm(formula = bpT ~ sex + obese, data = bp.obese)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.744e-05 -2.220e-05 -1.150e-07 1.802e-05 7.736e-05   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.880e-01 1.685e-05 34893.288 < 2e-16 \*\*\*  
## sex -1.806e-05 7.005e-06 -2.577 0.0114 \*   
## obese 6.085e-05 1.352e-05 4.500 1.86e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.204e-05 on 99 degrees of freedom  
## Multiple R-squared: 0.1745, Adjusted R-squared: 0.1579   
## F-statistic: 10.47 on 2 and 99 DF, p-value: 7.528e-05

plot(fit4)



AIC(fit3, fit4)

## df AIC  
## fit3 5 874.0163  
## fit4 4 -1816.6503

BIC(fit3, fit4)

## df BIC  
## fit3 5 887.1412  
## fit4 4 -1806.1504

La trasformazione ha permesso di rientrare nell’ipotesi di normalità dei residui e sia il test F del modello, sia le statistiche AIC e BIC confermano il fatto che fit4 sia il modello migliore.

Tuttavia il basso valore degli indici suggerisce il fatto che sono necessari ulteriori regressori per giungere a un modello soddisfacente.